Social and Personality Psychology Compass 5/10 (2011): 775–791, 10.1111/j.1751-9004.2011.00389.x

Data-driven Methods for Modeling Social Perception

Alexander Todorov[1,2]*, Ron Dotsch[1,2], Daniel H. J. Wigboldus[2], and Chris P. Said[3]

[1] *Princeton University*
[2] *Radboud University Nijmegen*
[3] *New York University*

Abstract

How do we model the complexity of social perception? A major methodological problem is that the space of possible variables driving social perceptions is infinitely large, thus posing an insurmountable hurdle for conventional approaches. Here, we describe a set of data-driven methods whose objective is to identify quantitative relationships between high-dimensional variables (e.g., visual images) and behaviors (e.g., perceptual decisions) with as little bias as possible. We focus on social perception of faces, although the methods could be applied to other visual and nonvisual categories. We review two reverse correlation approaches: (a) psychophysical methods based on judgments of images altered with randomly generated noise, where the analysis relates the random variations of the images to judgments; and (b) methods based on judgments of randomly generated faces from a statistical, multidimensional face space model, where the analysis relates the dimensions of the face model to judgments.

People are extremely adept at perceiving other people despite the complexity of social perception. In the case of face recognition, decades of computer science research have yet to produce a computer model that approximates human performance (Bowyer, Chang, & Flynn, 2006; Sinha, Balas, Ostrovsky, & Russell, 2006). After extremely brief exposure or highly degraded visual input, people can identify faces (Grill-Spector & Kanwisher, 2005; Yip & Sinha, 2002), their race and gender (Cloutier, Mason, & Macrae, 2005; Martin & Macrae, 2007), recognize their emotional expressions (Esteves & Öhman, 1993), and make a variety of social judgments such as aggressiveness (Bar, Neta, & Linz, 2006), trustworthiness (Todorov, Pakrashi, & Oosterhof, 2009), and sexual orientation (Rule & Ambady, 2008). People can also recognize familiar faces after more than 50 years (Bahrick, Bahrick, & Wittlinger, 1975).

How do people manage these perceptual and cognitive feats? Specifically, what perceptual information are they using to make inferences about social attributes? In this paper, we describe a set of data-driven methods that could be used to answer these questions. We focus on social perception of faces, although the methods that we describe could be applied to body perception, dynamic motion perception, and to nonvisual and multisensory modalities.

The face is the primary source of visual information for identifying people and for reading their emotional and mental states. People, with the exception of prosopagnosics who are unable to recognize faces and those suffering from disorders of social cognition such as autism, are extremely adept at these two tasks. Given that people agree in their social perception of faces (Todorov, Said, Engell, & Oosterhof, 2008; Zebrowitz & Montepare, 2008), it should be possible to model this perception. Specifically,

what differences in facial structure lead to appearance-based, social inferences? For example, what information do people use to decide that a face looks trustworthy or untrustworthy?

A major methodological problem is that the space of possible variables driving social perceptions is infinitely large, thus posing an insurmountable hurdle for conventional approaches. The standard approach is to systematically manipulate a facial feature (e.g., the corners of the mouth) and ask people to make social judgments of the face (e.g., friendliness). In most cases, these judgments would systematically change as a function of the changes in the feature. However, this finding does not necessarily show that the manipulated feature is the most important feature used in the judgments. First, changes of other features (e.g., the shape of the eyebrows) could lead to similar systematic changes in judgments. Second, the same feature would be perceived differently in the context of other features. Faces are perceived holistically as integrated gestalts rather than as a collection of independent features (Farah, Wilson, Drain, & Tanaka, 1998; Maurer, Le Grand, & Mondloch, 2002; Todorov, Loehr, & Oosterhof, 2010). Moreover, it is not even clear how features should be defined. For example, a mouth could be described as a feature but so could its parts (e.g., upper lip, lower lip, corners, etc.). Further, even with a relatively small number of features, the various feature combinations rapidly proliferate. With just 10 features, each one having only two possible values, there are 1024 different feature combinations. With 15 binary features, there are 32,768 combinations. With 20 binary features, there are 1,048,576 combinations. Clearly, traditional factorial experiments would hardly do the trick.

Here, we describe an alternative set of data-driven methods that combine insights from psychophysics, computer science, and experimental psychology. The objective of these methods is to identify a quantitative relationship between a high-dimensional variable (e.g., a visual face image) and a behavior (e.g., a perceptual decision) with as little bias as possible. In this respect, these are stimulus-driven techniques: although they are naturally constrained by the type of judgment and stimuli used in the specific experimental paradigm, they are not constrained by hypotheses that dictate specific manipulations of the images. We review two sets of techniques: psychophysical reverse correlation methods (PRCM; Gosselin & Schyns, 2004) and reverse correlation methods in the context of face space models (FSRCM; O'Toole, Wenger, & Townsend, 2001; Valentine, 1991). PRCM are based on judgments of images that are visually degraded or altered with randomly generated noise (Figure 1). FSRCM are based on judgments of randomly generated faces from a multidimensional face space model (Figure 2). In both techniques, the goal is to identify systematic relationships between stimulus parameters and social judgments.

Psychophysical Reverse Correlation Methods

These methods were originally developed in the domain of auditory cognition (Ahumada & Lovell, 1971), before they were used in research on vision (Ahumada, 2002; Solomon, 2002) and neurophysiology (Ringach & Shapley, 2004; Victor, 2005). The term 'reverse' refers to a reversal of the statistical relationship between stimulus and response. In conventional paradigms, responses depend on meaningful manipulation of stimulus attributes. This relationship is quantified by correlating fixed stimulus attributes with responses. In reverse correlation paradigms, on the other hand, variations in stimulus attributes are random. The correlation between stimuli and responses can be used to model those variations in stimulus attributes that caused the acquired response pattern. In this type of

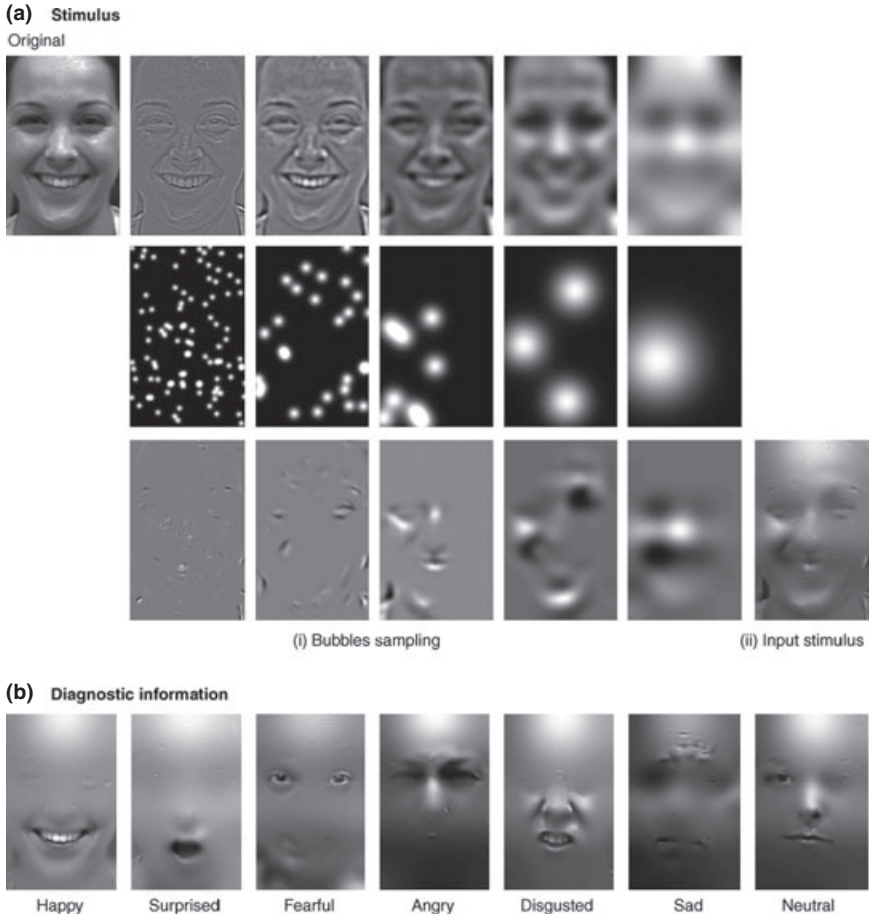

Figure 1 Illustrating PRCM. (a) Stimulus. (i) Bubbles sampling. In this application of Bubbles, an original face picture is initially filtered into five independent spatial frequency bands. On each sampling trial, Gaussian apertures are randomly allocated across spatial frequency bands and image locations to sample and reveal different portions of the input face. (ii) Input stimulus. On each trial, the input face will present different samples of facial information. As the location of the bubbles randomly change across trials, the entire face will be sampled throughout the experiment. (b) Diagnostic information. After many trials, a multiple linear regression associates correct and incorrect categorization responses (here, for different facial expressions) with the sampled facial features. Diagnostic information, therefore, represents the facial information that the observer's brain must process to correctly perform the task. In the example, different features are processed for correctly categorizing different facial expressions of emotions. Reprinted from Schyns et al. (2009), with permission from the authors and the publisher.

analysis, the response variable is fixed whereas the stimulus attributes are random. Because the response is caused by the sensory input and used as a basis of the classification of this input, the methods are called reverse correlation methods.

Applications to social perception

When it comes to social perception, two PRCM techniques have become popular. In both techniques, random noise is superimposed on visual images. However, whereas in the first technique – referred to as Bubbles (Gosselin & Schyns, 2001) – the underlying

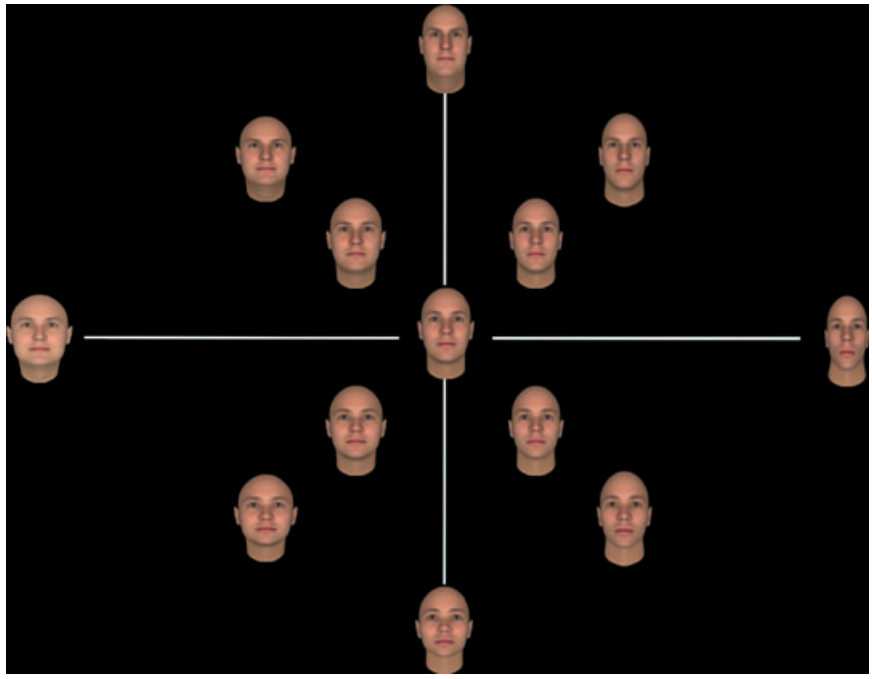

Figure 2 Illustrating the idea of a face space model using a simple two-dimensional space. The dimensions are extracted empirically from statistical analysis of 3D laser scans of real faces and define abstract, global properties of the faces. Here, for example, the first dimension on the *x*-axis is related to the width of the face. The second dimension on the *y*-axis is less easily described. The faces at the extremes of the dimensions (4 *SD*) from the average face exemplify the changes along the respective dimensions. The faces in the different quadrants are linear combinations of the two dimensions with the following coordinates: 1st quadrant (1, 1) and (2, 2), 2nd quadrant (1, −1) and (2, −2), 3rd quadrant (-1, -1) and (-2, -2), and 4th quadrant (-1, 1) and (-2, 2). In principle, one can generate an infinite number of faces within the plane defined by the dimensions.

image is unambiguous (e.g., a prototypical sad face), in the second technique – sometimes referred to as superstitious perception (Gosselin & Schyns, 2003; Mangini & Biederman, 2004) – the image is ambiguous (e.g., a morph of two facial expressions). Whereas the objective of the first technique is to reveal the diagnostic information used by the perceiver for the specific judgment (e.g., sadness), the objective of the second technique is to infer the perceiver's internal representation of the perceptual category (e.g., expression of sadness). We refer to these techniques as informational diagnosticity PRCM and internal representation PRCM, respectively.

With the informational diagnosticity PRCM technique, a researcher selects two images that are good exemplars of the categories of interest, for instance happy and sad expressions. In a single trial, one of the two faces is presented and participants are asked to classify the image into one of the categories: sad versus happy expression. Importantly, on every trial the presented face is degraded by superimposing a randomly generated mask, which reveals only small parts of the original image (see Figure 1). If a participant consistently makes a correct classification when specific parts of the face are unmasked, those parts are inferred to have diagnostic value. Reverse correlation analysis, in this case, identifies the visual areas of faces that are potentially informative for classification. This is conceptually different from eye-tracking studies, which identify the parts of the image the subject foveates to when presented with the full image.

The informational diagnosticity PRCM technique relies on the selection of a predefined signal (e.g., comparing happy and sad faces). In contrast, the internal representation PRCM technique is used when signal attributes are unknown or when researchers want to examine participants' subjective internal representation of a category, without making any assumptions about what typical category members look like. In a typical internal representation PRCM task, participants classify variations of one single base face that are unrelated to the categories of interest. These variations are created by distorting the base face with superimposed random noise (see Figure 3). Participants then classify the noisy faces in which ever categories interest the researchers. Averaging the noise patterns that were classified as belonging to the category of interest yields a classification image, showing a visual approximation of participants' subjective internal representation of the category.

In their validation of the internal representation PRCM technique, Mangini and Biederman (2004) demonstrated that the method works well for gender, emotional expression, and identity classification. For instance, in their Study 3, participants were asked to make identity judgments. Specifically, the base image was a morph between the faces of John Travolta and Tom Cruise. Random sinusoid noise was superimposed on the morph to create variations. Participants judged whether each variation was probably John Travolta, possibly John Travolta, possibly Tom Cruise, or probably Tom Cruise. Averaging all noise patterns classified as probably John Travolta resulted in a classification image (see Figure 4) showing what visual information yielded a John Travolta classification. Superimposing the classification image on the original base face resulted in an actual picture of John Travolta's face, or at the very least, an approximation of participants' subjective internal representation of his face. Likewise, averaging all noise patterns classified as probably Tom Cruise resulted in the Tom Cruise classification image.

Internal representation PRCM have only just begun to gain traction in basic social perception research. Dotsch, Wigboldus, Langner, and van Knippenberg (2008) used these methods to reveal potential biases in the representations of a stigmatized ethnic outgroup. In the Netherlands, Moroccans are a highly stigmatized immigrant group and are strongly associated with criminality (Gordijn, Koomen, & Stapel, 2001). In the experiment, Dutch participants were asked to repeatedly select the face that appeared most Morrocan from a pair of two noisy images presented side-by-side. As in the original Mangini and Biederman (2004) task, all stimuli were variations of the same base face. For example, in a single trial one stimulus consisted of the base image with superimposed random sinusoid noise, and the other of the same base image with the inverse noise superimposed. Averaging all stimuli selected as most Moroccan-looking yielded a Moroccan

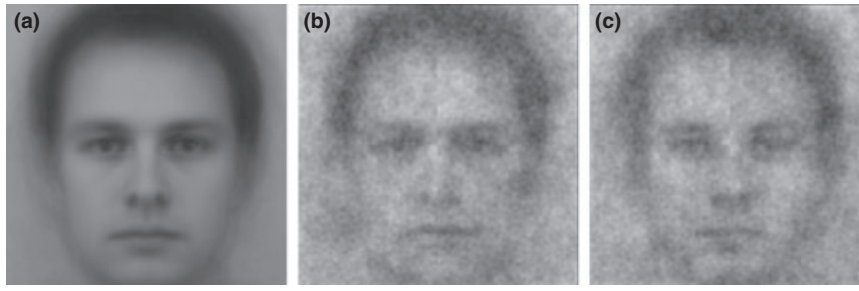

Figure 3 Illustrating internal representation PRCM. Example base face (a), base face with noise superimposed (b) and base face with inverse noise superimposed (c).

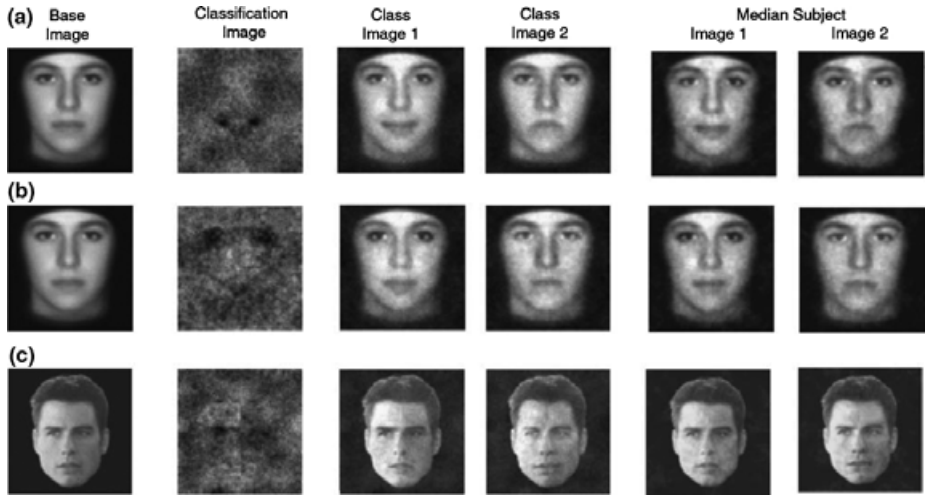

Figure 4 Results from three tasks using internal representation PRCM. The results are based on high confidence 'probably' responses (a) Happy/unhappy, (b) male/female, (c) Tom Cruise/John Travolta. The base images were identical for the expression and gender tasks. The darkest and lightest areas of the classification images indicate the areas that most influenced the participants' classifications. The addition of the classification image to the base face results in Class Image 1, which appears happy. The subtraction of the classification image results in Class Image 2, which appears unhappy. The same addition and subtraction operations produce the class images for (b) female and male and (c) Cruise and Travolta, respectively. The rightmost two columns show the classification images for the median subject calculated in terms of Euclidean pixel distance for a given subject's classification image and the average classification image. Reprinted from Mangini and Biederman (2004), with permission from the authors and the publisher.

classification image (Figure 5), which represented what participants thought a typical Moroccan face looks like. Importantly, more implicitly prejudiced participants (as measured with a single target implicit association test, see Bluemke & Friese, 2008; Greenwald, McGhee, & Schwartz, 1998) generated classification images that looked more criminal and less trustworthy compared to the classification images of the less prejudiced participants.

These methods could be used to reveal the internal representations of any social category. Dotsch and Todorov (under review) applied the methods to judgments of trustworthiness, dominance, and threat. As in the Dotsch et al. (2008) study, participants were presented with a pair of identical base images that differed only in the random noise superimposed on them. Averaging across the images selected to be trustworthy revealed

Figure 5 Moroccan classification image resulting from internal representation PRCM (Dotsch et al., 2008).

the representation of trustworthy faces. Similarly, averaging across the images selected to be untrustworthy revealed the representation untrustworthy faces. The same procedures were applied for judgments of dominance and threat (Figure 6). Importantly, the visual features identified in these representations were similar to the features obtained with FSRCM, methods that we consider below.

In the context of face perception, these methods have been applied to perception of gender (Nestor & Tarr, 2008), identity (Mangini & Biederman, 2004), expressions of emotions (Jack, Caldara, & Schyns, 2011; Langner, Becker, & Rinck, 2009; Smith, Cottrell, Gosselin, & Schyns, 2005), groups (Dotsch, Wigboldus, & van Knippenberg, 2011; Dotsch et al., 2008; Imhoff, Dotsch, Bianchi, Banse, & Wigboldus, forthcoming), personality traits (Dotsch et al., under review), and Mona Lisa's smile (Kontsevich & Tyler, 2004). These methods could also be employed for topics as diverse as attractiveness and stereotype formation. However, potential research topics need not be confined to face perception. Visual perception of objects other than faces might also be reverse correlated. Using noisy sound signals, auditory perception can be investigated with PRCM (coming back to where the method actually originated; Ahumada & Lovell, 1971). Biological motion, such as human gait patterns, could in principle be reverse correlated (using decomposition methods such as those suggested by Troje, 2002). Indeed, the possibilities of reverse correlation methods are endless, limited only by the fact that researchers should be able to generate random variations of stimuli. PRCM accomplish this using noise. FSRCM, which are discussed next, achieve this in a different way: by varying properties of the faces directly.

Reverse Correlation Methods in the Context of Face Space Models

There are two basic tasks in this approach: creating a statistical model of face representation and using this model to derive the changes in facial features that lead to corresponding changes in social judgments. Before we describe the specific technique and its applications to social perception, we briefly describe the idea of face space, first proposed by Valentine (1991). According to this idea, faces are represented as points in a multidimensional space, where each dimension is a property of the face. Valentine used this idea to account for a number of face recognition findings, including effects of distinctiveness (recognition advantage for distinctive faces) and race (recognition advantage for own race faces). Subsequently, Blanz and Vetter (1999, 2003) developed realistic face models based on 3D laser scans of real faces[1]. O'Toole and her colleagues (O'Toole, Vetter, & Blanz, 1999; O'Toole, Vetter, Troje, & Bülthoff, 1997) used these models to test a number of interesting hypotheses, including the contribution of shape and reflectance information to face recognition. These multidimensional models provide a powerful representational

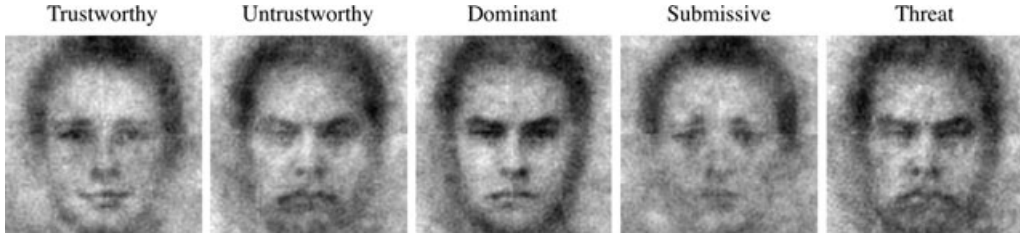

Figure 6 Trustworthy, untrustworthy, dominant, submissive, and threat classification images resulting from internal representation PRCM.

framework that can account for variations in face identity and facial expressions (Calder & Young, 2005).

To illustrate the idea of a multidimensional face space, imagine that two dimensions are sufficient to explain variation in faces (Figure 2), and that the first dimension is related to the width of the face and the second dimension is related to the height of the face. Each face positioned on this 2D plane will have coordinates for each of the dimensions. With this model, one can generate an infinite number of faces within the plane defined by the two dimensions by changing the face coordinates on the two dimensions. In reality, more dimensions are needed to represent the natural variation in human faces, and many studies have used a face space containing 50 dimensions or more. These dimensions are empirically derived from natural variation in faces (Blanz & Vetter, 1999, 2003). In some cases, faces of real people are laser scanned in 3D, a surface mesh with defined topology is imposed on each 3D image, and then the images are analyzed with statistical techniques that exploit the correlations between the vertices that define the 3D images

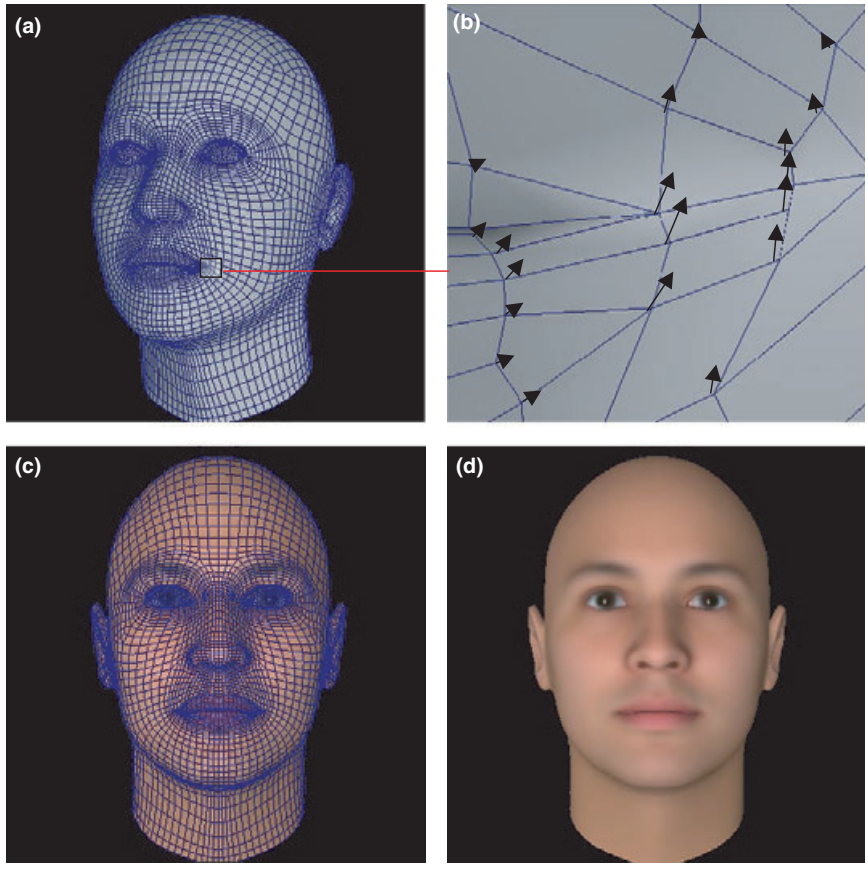

Figure 7 Illustration of how statistical face models represent faces. (a) Side view showing a surface mesh superimposed on the average face. (b) Linear changes in the vertex positions for the surface of a face segment on one of the shape dimensions. (c) Frontal view of the average face showing the surface mesh and texture. (d) Frontal view of the average face with texture. Face shape for specific faces is represented as a function of the average face and the differences from this face for all vertices. These differences are captured in the principal components for shape. Reprinted from Oosterhof and Todorov (2008), *PNAS*, Copyright 2008 National Academy of Sciences, U.S.A.

(Figure 7). The final result is a statistical model of face representation defined by a fixed number of dimensions. This model can generate an unlimited number of faces, where each face is a point defined by its coordinates on all face dimensions.

Applications to social perception

With the aid of a statistical face model, it is relatively straightforward to uncover the variations in the structure of faces that lead to specific face evaluations. As explained above, in these models each face is numerically represented by a set of values on the model dimensions. These dimensions do not need to be psychologically meaningful. Instead, FSRCM determine how much each dimension contributes to psychologically meaningful judgments (e.g., perceived face trustworthiness). Then, it is possible to use these weights to construct new dimensions, along which the psychologically meaningful judgment varies maximally.

To conceptually understand this process, imagine a set of faces that are rated on a social attribute. Each face will have a score on this attribute and the values on the visual dimensions that define the face space. The correlations between the attribute measure and the face coordinates determine the direction of the attribute dimension in the face space. For example, if the attribute measure is highly correlated with one visual dimension, then large changes along the attribute dimension will correspond to large changes along this visual dimension. If the attribute measure is weakly correlated with another visual dimension, then large changes along the attribute dimension will correspond to small changes along this visual dimension. The resulting attribute dimension is a linear combination of the visual dimensions defining the face space. Under most circumstances, the correlation approach provides a very close estimate to the true (maximum likelihood) linear relationship between the dimensions and the social attribute. A more exact approach is to use linear regression. Under this approach, the attribute values are regressed on the face dimensions values. The set of regression weights then defines the direction in face space along which the social attribute is predicted to vary maximally.

Similar to PRCM, there is no explicit manipulation of facial features in FSRCM. The derived psychological dimensions are defined by social judgments of unmanipulated faces rather than by the ideas of the experimenter. To derive a meaningful dimension in the statistical face space, one first needs to randomly generate a large sample of faces, using the statistical model. Second, these faces are rated on social dimensions. Third, the ratings provide the input for statistical analysis that specifies the social dimension. Fourth, the new social dimension can be applied to novel faces (Oosterhof & Todorov, 2008). This approach can be used to model face perception for any arbitrarily chosen social dimension (e.g., extroversion, risk-seeking, emotional stability, etc.). Oosterhof and Todorov (2008) used these methods to build models of perceived face trustworthiness, dominance, and threat (Figure 8). Todorov and Oosterhof (2011) provided additional information about these methods and built models of several other social dimensions, including attractiveness (see also Said & Todorov, forthcoming). Walker and Vetter (2009) built models of six different social dimensions: aggressiveness, extroversion, likeability, risk-seeking, social skills, and trustworthiness (Figure 9). Moreover, they applied these models to faces of real people and showed that subtle manipulations in their pictures resulted in the expected social attributions.

These methods discover structural differences in appearance that predict differences in social perception. For example, Figure 10 shows face variations on five social dimensions – trustworthiness, dominance, threat, likeability, and competence – that were derived

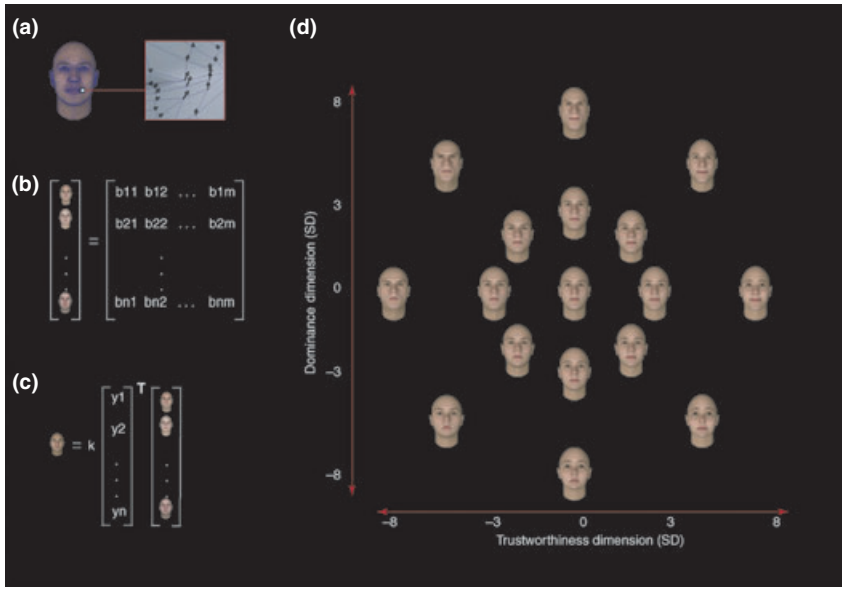

Figure 8 Application of FSRCM to social judgments of faces. (a) Illustration of how the face model represents faces. Left: A surface mesh with fixed topology superimposed on the average face. Right: an expanded view of a section of the mesh, along with direction vectors specifying the linear changes in the vertex positions for the surface for one of the $m = 50$ shape dimensions. (b) A set of n random faces can be obtained by linear combinations of the m shape components, and represented in an n by m matrix. These dimensions are extracted from a principal component analysis of shape variations of the vertex positions and do not necessarily have inherent psychological meaning. Each row of the matrix contains the set of m weighting coefficients corresponding to a particular face. (c) Each of the n faces is rated by participants on a trait dimension and given an average score y_i. Multiplication of the social judgments vector by the set of randomly generated faces yields a dimension that is optimal in changing faces on the trait dimension, which can be controlled with a tunable constant k. The figure shows the generation of one face along the trustworthiness dimension. (d) A two-dimensional model of evaluation of faces. Examples of a face with exaggerated features on the two orthogonal dimensions – trustworthiness plotted on the x-axis and dominance plotted on the y-axis – of face evaluation. The changes in features were implemented in a computer model based on trustworthiness and dominance judgments of $n = 300$ emotionally neutral faces (Oosterhof & Todorov, 2008). The extent of face exaggeration is presented in SD units. The faces on the diagonals were obtained by averaging the faces on the trustworthiness and dominance dimensions. The diagonal dimension passing from the 2nd to the 4th quadrant was nearly identical to a dimension based on threat judgments of faces. The other diagonal dimension passing from the 1st to the 3rd quadrant was similar to dimensions empirically obtained from judgments of likeability, extraversion, and competence. Reprinted from Todorov et al. (2008), with permission from the authors and the publisher.

with FSRCM (see Todorov & Oosterhof, 2011 for details). The first row of the figure shows five versions of a face manipulated to decrease (the two faces to the left of the center face) or increase its trustworthiness (the two faces to the right). As the predicted trustworthiness ratings of the face increase, the face appears to change its expression from neutral to happy. In contrast, as the predicted trustworthiness ratings of the face decrease, the face appears to change its expression from neutral to angry. In the case of dominance (second row of Figure 10), as face dominance increases, the face appears more masculine and mature-faced. Face threat (third row of Figure 10) could be reproduced as a combination of face untrustworthiness and face dominance. The most threatening face is a dominant face that appears angry. It should be noted that these findings nicely converge with findings from PRCM studies (Figure 6). In the case of face likeability and face competence (fourth and fifth rows of Figure 10), as the perceived likeability and competence of the face increases, the face becomes more attractive and mature-faced.

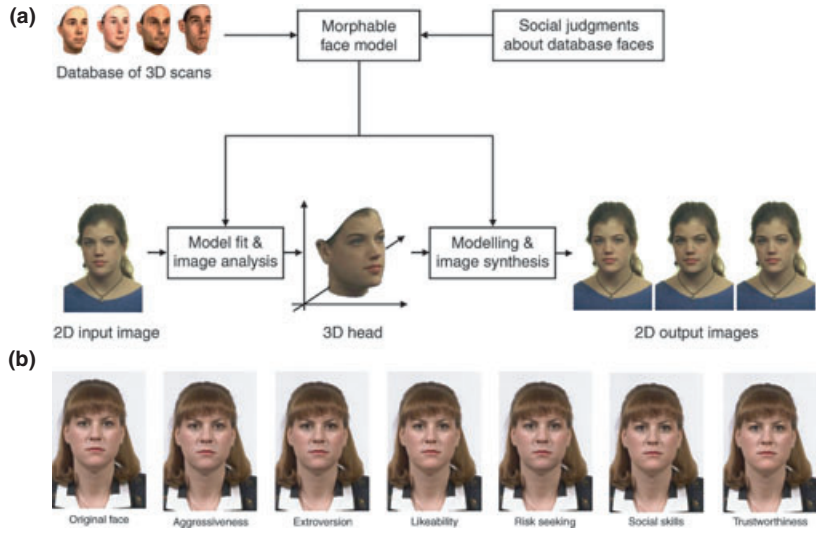

Figure 9 Application of FSRCM to social judgments of faces. (a) Using a morphable face model based on 200 3D scans of faces, it is possible to locate the directions in face space with maximum variability for specific social judgments. Any photograph of any human face can be analyzed by fitting this model to the face image. The resulting 3D head can be manipulated by adding or subtracting personality trait vectors and rendered back into the original photograph. (b) Examples of applying six different personality trait vectors to a natural face image. Reprinted from Walker and Vetter (2009), with permission from the authors and the publisher.

By exaggerating the features that contribute to social judgments of emotionally neutral faces, it is possible to reveal the underlying variations that account for these judgments. One way to think about this process is in terms of creating a caricature of a face on the dimension of interest or in terms of amplifying the diagnostic signal in the face that is used for the specific judgment. For example, in the case of trustworthiness, although faces are perceived as emotionally neutral within the range shown in Figure 10, they are perceived as emotionally expressive outside of this range (see Figure 8d). Specifically, whereas faces at the extreme negative end of the dimension appear to express anger, faces at the extreme positive end appear to express happiness. This finding is consistent with theories that posit that social inferences from facial appearance are based on resemblance to features that have adaptive significance (Zebrowitz & Montepare, 2008). For example, expressions of emotions are indicative of mental and emotional states and provide signals for appropriate behaviors. As a result, resemblance of faces to specific emotional expressions can be misattributed to stable personality characteristics (Said et al., 2009).

Limitations

The data-driven methods described here provide powerful tools for identifying diagnostic visual information for particular tasks (as in the informational diagnosticity PRCM and the FSRCM) and for identifying internal representations of particular categories (as in the internal representation PRCM and the FSRCM). However, as with any method, these methods have some inherent limitations.

First, as Mangini and Biederman (2004) have noted, it is very tempting to equate the outcomes of reverse correlation paradigms to actual mental representations. However, the

Figure 10 Faces generated by computer models of social perception. These models were empirically derived from social judgments of a large sample of randomly generated faces and then applied to the average face presented in the third column of faces in the figure. The perceived value of the faces on the respective social dimensions increases from left to right.

outcome is a quantification of the strategy used when performing the task, which to a large extent correlates with mental representations. For instance, the outcome of PRCM is a visual classification image from which researchers may infer underlying mental representations. The classification image itself, however, is a function not only of these mental representations, but also of the specific base image and noise patterns used, as well as the motivation of a participant to perform the task. This may be more of concern to PRCM – where participants are asked to judge a very large number of artificially degraded stimuli – than to FSRCM where the stimuli are nondegraded images of faces that approximate natural social perception. Nevertheless, for both methods, the outcome in principle may not reveal anything about the cognitive properties of underlying mental representations. These representations could be prototypes, exemplars, feature lists, or abstract rules that could potentially yield the same reverse correlation outcome. Further, it is important to keep in mind that identifying diagnostic features in a visual image does not directly translate into identifying actual perception mechanisms. For example, PRCM require a large number of trials that by virtue of the noise masking procedures induce a specific form of perceptual processing that is not necessarily representative of perception in many everyday tasks.

Second, not every topic lends itself well for reverse correlation application. Due to the number of trials (relative to the large number of stimulus attribute variations) reverse correlation data are commonly analyzed in a linear fashion and interactions between features are mostly disregarded.[2] This can be problematic depending on what is being reverse

correlated. For instance, Mangini and Biederman (2004) give the example of reverse correlating an eyewink. Participants would classify faces with a closed eye as showing a wink, regardless of whether it is the left or the right eye. The outcome, which is the average of these faces, would show a face with both eyes half-closed.

Third, although the number of trials in reverse correlation paradigms is typically too small to model interactions, it is often too large for participants to stay motivated until the end of a task. It is not uncommon for a reverse correlation task to consist of hundreds or even thousands of trials. When participants respond randomly due to the high number of trials, the quality of the procedure is degraded severely. The large number of trials may be more of concern to PRCM than to FSRCM. However, even the latter methods require judgments of a large number of faces. For example, creating a social dimension in a 50-dimensional face space model requires fitting 50 parameters. Hence, the researcher would need judgments of, at least, several hundred randomly generated faces.

Finally, the specific set of stimuli used in reverse correlation tasks may pose a concern to external validity. In PRCM, participants often classify far less stimuli than the number of randomly varied stimulus dimensions. As a consequence, it is possible that resulting classification images are affected to a large extent by the specific set of used stimuli rather than by the underlying psychological representations. Increasing the number of trials, or creating unique random stimulus sets for each participant may enhance external validity.

Similarly, because stimuli in FSRCM tasks are generated using a pre-defined face space, the method of constructing that face space may influence reverse correlation results. For instance, a face space derived from a limited source set of Caucasian faces can be used to generate predominantly Caucasian-looking stimuli, whereas African American faces will be harder to generate (although interpolation and extrapolation enable the space to represent faces that were not originally part of the source set). Thus, the randomly generated stimulus faces in FSRCM, and therefore the reverse correlation results, are constrained by the ethnicity, age, gender and other properties of the source faces used to build the face space.

Future Method Development and Research Applications

Over time, reverse correlation methods may be improved in order to generate higher quality outcomes or reduce the number of trials. One promising development in this respect is the implementation of adaptive search algorithms, which base the presentation of stimuli later in the task on participants' responses earlier in the task. Another new development is the use of multiple response alternatives (Dai & Micheyl, 2010). Reverse correlation methods have been limited to two categories (e.g., Tom Cruise versus John Travolta, happy versus sad) or one category or dimension (e.g., Moroccan, trustworthy, dominant, etc.) per task. Outcomes may change considerably when a participant simultaneously considers multiple categories or dimensions. For example, some of the faces classified by participants as Moroccan might also be classified as Turkish if participants are given this alternative. Multiple response alternatives reverse correlation makes it possible to disentangle multiple categories or dimensions by allowing participants to classify stimuli across a set of categories.

Reverse correlation methods are particularly useful for identifying commonalities and differences in information use in different populations. As such, these methods can be used to construct culture-specific models of social judgments from the ground up and, hence, address important questions about the universal or culture-specific nature of social perception. As described above, the methods for constructing such models are data-driven

without imposing a priori assumptions about the importance of specific facial features (e.g., nose, eye brows, etc.). All that is needed is social judgments of faces by samples representative for specific cultural groups. In fact, such methods could be used for comparative, cross-species research. For example, Martin-Malivel, Mangini, Fagot, and Biederman(2006) used PRCM to study differences between humans and baboons in information use in a face classification task.

The same methods can be applied to build models specific to individuals and use these models to explore individual differences in social perception. These methods can be easily extended to study people with abnormal social cognition including autism and schizophrenia. In fact, Langner et al. (2009) used PRCM to study differences between socially anxious and non-anxious participants in discriminating emotional expressions.

FSRCM methods could be used to build 3D models of social perception. Then, face avatars generated by these models could be used in immersive virtual reality environments to study social interaction. For example, one can experimentally study the consequences of first impressions in life like social interactions. Finally, these faces could be useful in research and design of human computer interfaces. Using empirically derived models of social perception, it should be possible to create avatars that would be most appropriate for specific communications (e.g., providing advice vs. threat warning).

To conclude, we think that the methods described here hold great promise for revealing the perceptual basis of social judgments and we hope that these methods would reach a wider use among scientists interested in social perception.

Acknowledgement

This research was supported by National Science Foundation grant 0823749 and the Russell Sage Foundation.

Short Biographies

Alexander Todorov is an Associate Professor of Psychology and Public Affairs at Princeton University with a joint appointment in the Department of Psychology and the Woodrow Wilson School of Public and International Affairs. He is also an affiliated faculty of the Princeton Neuroscience Institute and a visiting professor at Radboud University Nijmegen in the Netherlands. His research focuses on the cognitive and neural basis of social cognition. His main line of research is on the cognitive and neural mechanisms of person perception with a particular emphasis on the social dimensions of face perception. His research approach is multidisciplinary, using a variety of methods from behavioral and functional Magnetic Resonance Imaging experiments to computer and statistical modeling. Alexander received his PhD in Psychology from New York University in 2002.

Ron Dotsch is a postdoctoral researcher at the Department of Psychology at Princeton University. He is currently working with Alexander Todorov to study social face perception. He received his PhD in Psychology from Radboud University Nijmegen, where he worked with Daniel Wigboldus and Ad van Knippenberg. Ron's research interests include face perception, social categorization, prejudice, stereotypes, and the use of advanced methods, such as virtual reality, reverse correlation, and face space models.

Daniel Wigboldus is a Professor of Social Psychology at the Behavioural Science Institute of the Radboud University Nijmegen in the Netherlands. His research interests

concerns person perception in general and stereotyping and prejudice in particular. He is a behavioral scientist that likes to use multiple methods in his research varying from language use analysis to immersive virtual environment technology. His current research focuses on how prejudice and stereotypes affect face perception. Daniel received his PhD in Psychology from the Free University, Amsterdam in the Netherlands.

Christopher Said is a postdoctoral research scientist at the Department of Psychology and the Center for Neural Science at New York University. He is currently working with David Heeger to study computational models of visual attention using fMRI and EEG. He received his doctorate in psychology and neuroscience from Princeton University, where he worked with Alexander Todorov. Christopher is interested in the effects of attention on early visual cortex. He is also interested in using face space models and regression approaches to infer the properties of social judgments from faces.

Endnotes

* Correspondence address: Green Hall, Princeton, NJ 08540, USA. Email: atodorov@princeton.edu

[1] We focus on these 3D models but the very first computational face models were based on principal components analysis of the pixel intensities of 2D facial images (Turk & Pentland, 1991). Using this approach, to the best of our knowledge, Brahnam (2005) was the first to build a computational model of trait impressions from faces. Perhaps because this work was published in a journal that ceased to exist after its first volume, the research is not well known.

[2] In the introduction, we pointed out that one of the weaknesses of standard approaches is the impossibility to model all feature interactions. The difference between these approaches and data-driven approaches described here is that the latter allow for the stimuli to vary across the whole face (and therefore all possible features) without limiting the search to specific features. This makes it possible for solutions to emerge that not only show the effects of specific feature but also effects of interacting features on social perception.

References

Ahumada, A. J. (2002). Classification image weights and internal noise level estimation. *Journal of Vision*, **2**, 121–131.

Ahumada, A. J., & Lovell, J. (1971). Stimulus features in signal detection. *The Journal of the Acoustical Society of America*, **49**, 1751–1756.

Bahrick, H. P., Bahrick, P. O., & Wittlinger, R. P. (1975). Fifty years of memory for names and faces: A cross-sectional approach. *Journal of Experimental Psychology: General*, **104**, 54–75.

Bar, M., Neta, M., & Linz, H. (2006). Very first impressions. *Emotion*, **6**, 269–278.

Blanz, V., & Vetter, T. (1999). A morphable model for the synthesis of 3D faces. Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques SIGGRAPH 99, **33**, Annual Conference Series, ACM Press, pp. 187–194. ISSN: 00978930, ISBN: 0201485605. doi: 10.1145/311535.311556.

Blanz, V., & Vetter, T. (2003). Face recognition based on fitting a 3D morphable model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **25**, 1063–1074.

Bluemke, M., & Friese, M. (2008). Reliability and validity of the Single-Target IAT (ST-IAT): Assessing automatic affect toward multiple political groups. *European Journal of Social Psychology*, **38**, 977–997.

Bowyer, K. W., Chang, K., & Flynn, P. (2006). A survey of approaches and challenges in 3D and multi-modal 3D + 2D face recognition. *Computer Vision and Image Understanding*, **101**, 1–15.

Brahnam, S. (2005). A computational model of the trait impressions of the face for agent perception and face synthesis. *AISB Journal*, **1**(6), 481–508.

Calder, A. J., & Young, A. W. (2005). Understanding the recognition of facial identity and facial expression. *Nature Reviews Neuroscience*, **6**, 641–651.

Cloutier, J., Mason, M. F., & Macrae, C. N. (2005). The perceptual determinants of person construal: Reopening the social-cognitive toolbox. *Journal of Personality and Social Psychology*, **88**, 885–894.

Dai, H., & Micheyl, C. (2010). Psychophysical reverse correlation with multiple response alternatives. *Journal of Experimental Psychology: Human Perception and Performance*, **36**(4), 976–993.

Dotsch, R., Wigboldus, D. H. J., & van Knippenberg, A. (2011). Biased allocation of faces to social categories. *Journal of Personality and Social Psychology*, **100**, 999–1014.

Dotsch, R., Wigboldus, D. H., Langner, O., & van Knippenberg, A. (2008). Ethnic out-group faces are biased in the prejudiced mind. *Psychological Science*, **19**, 978–980.
Esteves, F., & Öhman, A. (1993). Masking the face: Recognition of emotional facial expressions as a function of the parameters of backward masking. *Scandinavian Journal of Psychology*, **34**, 1–18.
Farah, M. J., Wilson, K. D., Drain, M., & Tanaka, J. N. (1998). What is ''special'' about face perception? *Psychological Review*, **105**, 482–498.
Gordijn, E. H., Koomen, W., & Stapel, D. A. (2001). Level of prejudice in relation to knowledge of cultural stereotypes. *Journal of Experimental Social Psychology*, **37**, 150–157.
Gosselin, F., & Schyns, P. G. (2001). Bubbles: A technique to reveal the use of information in recognition tasks. *Vision Research*, **41**, 2261–2271.
Gosselin, F., & Schyns, P. G. (2003). Superstitious perceptions reveal properties of internal representations. *Psychological Science*, **15**(5), 505–509.
Gosselin, F., & Schyns, P. G. (2004). (Eds.) Special issue: Rendering the use of visual information from spiking neurons to recognition. *Cognitive Science*, **28** (2), 141–146.
Greenwald, A. G., McGhee, D. E., & Schwartz, J. K. L. (1998). Measuring individual differences in implicit cognition: The Implicit Association Test. *Journal of Personality and Social Psychology*, **74**, 1464–1480.
Grill-Spector, K., & Kanwisher, N. (2005). Visual recognition: As soon as you know it is there, you know what it is. *Psychological Science*, **16**, 152–160.
Imhoff, R., Dotsch, R., Bianchi, M., Banse, R., & Wigboldus, D. H. J. (forthcoming). Facing Europe: Visualizing spontaneous in group projection. *Psychological Science*.
Jack, R. E., Caldara, R., & Schyns, P. G. (2011). Internal representations reveal cultural diversity in expectations of facial expressions of emotion. *Journal of Experimental Psychology: General*, doi: 10.1037/a0023463.
Kontsevich, L. L., & Tyler, C. W. (2004). What makes Mona Lisa smile? *Vision Research*, **44**, 1493–1498.
Langner, O., Becker, E. S., & Rinck, M. (2009). Social anxiety and anger identification: Bubbles reveal differential use of facial information with low spatial frequencies. *Psychological Science*, **20**(6), 666–670.
Mangini, M. C., & Biederman, I. (2004). Making the ineffable explicit: Estimating the information employed for face classification. *Cognitive Science*, **28**, 209–226.
Martin, D., & Macrae, C. N. (2007). A boy primed Sue: Feature based processing and person construal. *European Journal of Social Psychology*, **37**(5), 793–805.
Martin-Malivel, J., Mangini, M. C., Fagot, J., & Biederman, I. (2006). Do humans and baboons use the same information when categorizing human and baboon faces? *Psychological Science*, **17**, 599–607.
Maurer, D., Le Grand, R., & Mondloch, C. J. (2002). The many faces of configural processing. *Trends in Cognitive Sciences*, **6**, 255–260.
Nestor, A., & Tarr, M. J. (2008). Gender recognition of human faces using color. *Psychological Science*, **19**, 1242–1246.
Oosterhof, N. N., & Todorov, A. (2008). The functional basis of face evaluation. *Proceedings of the National Academy of Sciences of the USA*, **105**, 11087–11092.
O'Toole, A. J., Vetter, T., & Blanz, V. (1999). Three-dimensional shape and two-dimensional surface reflectance contributions to face recognition: An application of three-dimensional morphing. *Vision Research*, **39**, 3145–3155.
O'Toole, A. J., Vetter, T., Troje, N. F., & Bülthoff, H. H. (1997). Sex classification is better with three-dimensional head structure than with image intensity information. *Perception*, **26**, 75–84.
O'Toole, A. J., Wenger, M. J., & Townsend, J. T. (2001). Quantitative models of perceiving and remembering faces: Precedents and possibilities. In M. J. Wenger & J. T. Townsend (Eds.), *Computational, Geometric, and Process Perspectives on Facial Cognition* (pp. 1–38). Mahwah, New Jersey: Lawrence Erlbaum Associates.
Ringach, D. L., & Shapley, R. (2004). Reverse correlation in neurophysiology. *Cognitive Science*, **28**, 147–166.
Rule, N. O., & Ambady, N. (2008). Brief exposures: Male sexual orientation is accurately perceived at 50 ms. *Journal of Experimental Social Psychology*, **44**, 1100–1105.
Said, C., Sebe, N., & Todorov, A. (2009). Structural resemblance to emotional expressions predicts evaluation of emotionally neutral faces. *Emotion*, **9**, 260–264.
Said, C. P., & Todorov, A. (forthcoming). A statistical model of facial attractiveness. *Psychological Science*.
Schyns, P. G., Gosselin, F., & Smith, M. L. (2009). Information processing algorithms in the brain. *Trends in Cognitive Sciences*, **13**, 20–26.
Sinha, P., Balas, B., Ostrovsky, Y., & Russell, R. (2006). Face recognition by humans: Nineteen results all computer vision researchers should know about. *Proceedings of the IEEE*, **94**, 1948–1962.
Smith, M. L., Cottrell, G. W., Gosselin, F., & Schyns, P. G. (2005). Transmitting and decoding facial expressions. *Psychological Science*, **16**, 184–189.
Solomon, J. A. (2002). Noise reveals visual mechanisms of detection and discrimination. *Journal of Vision*, **2**, 105–120.
Todorov, A., Loehr, V., & Oosterhof, N. N. (2010). The obligatory nature of holistic processing of faces in social judgments. *Perception*, **39**, 514–532.

Todorov, A., & Oosterhof, N. N. (2011). Modeling social perception of faces. *Signal Processing Magazine, IEEE*, **28**, 117–122.
Todorov, A., Pakrashi, M., & Oosterhof, N. N. (2009). Evaluating faces on trustworthiness after minimal time exposure. *Social Cognition*, **27**, 813–833.
Todorov, A., Said, C. P., Engell, A. D., & Oosterhof, N. N. (2008). Understanding evaluation of faces on social dimensions. *Trends in Cognitive Sciences*, **12**, 455–460.
Troje, N. F. (2002). Decomposing biological motion: A framework for analysis and synthesis of human gait patterns. *Journal of Vision*, **2**, 371–387.
Turk, M., & Pentland, A. (1991). Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, **3**, 71–86.
Valentine, T. (1991). A unified account of the effects of distinctiveness, inversion, and race in face recognition. *The Quarterly Journal of Experimental Psychology*, **43A**(2), 161–204.
Victor, J. D. (2005). Analyzing receptive fields, classification images and functional images: Challenges with opportunities for synergy. *Nature Neuroscience*, **8**, 1651–1656.
Walker, M., & Vetter, T. (2009). Portraits made to measure: Manipulating social judgments about individuals with a statistical face model. *Journal of Vision*, **9**(11), 12.
Yip, A. W., & Sinha, P. (2002). Contribution of color to face recognition. *Perception*, **31**(8), 995–1003.
Zebrowitz, L. A., & Montepare, J. M. (2008). Social psychological face perception: Why appearance matters. *Social and Personality Psychology Compass*, **2**, 1497–1517.